ADVANCES IN MATHEMATICS 22, 364–385 (1976)

Developments in Heilbronn's Triangle Problem

K. F. Roth

Department of Mathematics, Imperial College of Science and Technology, London, SW7 5HH England

DEDICATED TO THE MEMORY OF MY GREAT FRIEND NORMAN LEVINSON

1. Introduction

Let $P_1, P_2, \ldots, P_n$ be a distribution of n points (where $n \geqslant 3$) in a disc of unit area. It was conjectured by Heilbronn, over 25 years ago, that the minimum of the areas of the triangles $P_iP_jP_k$ (taken over all selections P_i, P_j, P_k of three out of the n points) is less than cn^{-2}, where the constant c is absolute (and, in particular, independent of the distribution).

This conjecture remains open, although various partial results are known. We shall give a descriptive account of the methods that have been applied to the problem, endeavoring to explain their motivation and to make clear the various ideas that are involved. It seems probable that at least some of these ideas, perhaps modified and in new combinations, will feature in future progress.

2. Reformulation of the Problem

Let $\mathbf{K}$ be a finite closed convex region in the plane, of area $A(\mathbf{K})$. Let

$$P_1, P_2, \ldots, P_n \tag{2.1}$$

(where $n \geqslant 3$) be a distribution of n points in $\mathbf{K}$, such that the minimum of the areas of the triangles $P_iP_jP_k$ (taken over $1 \leqslant i < j < k \leqslant n$) assumes its maximum possible value $\triangle^*(\mathbf{K}; n)$. Heilbronn's problem, in the form stated in Section 1, is that of estimating the expression

$$\triangle(\mathbf{K}; n) = \triangle^*(\mathbf{K}; n)/A(\mathbf{K}) \tag{2.2}$$

(in terms of n) when $\mathbf{K}$ is a disc.

The corresponding problem for unrestricted **K** is only superficially more general. By considering linear transformations of the plane, we see that when **K** is a triangle (together with its interior), the value of (2.2) is independent of the particular triangle. Since any finite convex region **K** contains a triangle of area $\frac{1}{4}A(\mathbf{K})$ and lies inside a triangle of area $4A(\mathbf{K})$, it follows that the value of (2.2) is always at least a quarter and at most four times the value it assumes when **K** is a triangle. Thus the problem of estimating (2.2) is essentially (to within constant factors) independent of **K**. Although regions **K** other than discs have been considered, this is purely a matter of technical convenience; for example, use has been made of the circumstance that when **K** is a triangle the value of (2.2) is independent of the particular triangle, in conjunction with the fact that every polygon can be split into triangles.

We shall use

$$\triangle(\mathbf{D}; n), \qquad \triangle(\mathbf{S}; n), \qquad \triangle(\mathbf{T}; n) \tag{2.3}$$

to denote the values of (2.2) when **K** is, respectively, a disc, a square, a triangle; these quantities are of course independent of the particular disc, square or triangle. Heilbronn's problem may be interpreted to be that of estimating any one of these quantities when n is large.

We shall write simply $\triangle(n)$ in place of $\triangle(\mathbf{D}; n)$, and *henceforth we shall suppose that n is large.*

3. The Literature of the Problem

It was shown by Erdös (see [1, Appendix]) that

$$\triangle(n) \gg n^{-2}, \tag{3.1}$$

so that Heilbronn's conjecture, if true, would be best possible.

We note that by joining one of the points (2.1) (P_1 say) to the remaining $n - 1$ points we can form, in an obvious way, $n - 2$ triangles of type $P_1P_jP_k$ with disjoint interiors. Thus $\triangle(n) < 1/(n - 2)$, and the inequality

$$\triangle(n) \ll n^{-1} \tag{3.2}$$

is trivial. The first nontrivial estimate was due to Roth [1], who in 1950 proved that

$$\triangle(n) \ll n^{-1}(\log\log n)^{-1/2}. \tag{3.3}$$

There was no further improvement until about 20 years later, when Schmidt [2] established

$$\triangle(n) \ll n^{-1}(\log n)^{-1/2} \tag{3.4}$$

and indeed

$$\triangle(\mathbf{S}; n) < 500n^{-1}(\log n)^{-1/2}.$$

Subsequently Roth [3] obtained

$$\triangle(n) \ll n^{-\mu+\epsilon}, \qquad \textit{where} \quad \mu = 2 - (4/5)^{1/2} = 1.105\ldots,$$

and later [4] improved this to

$$\triangle(n) \ll n^{-\mu'+\epsilon}, \tag{3.5}$$

where

$$\mu' = \tfrac{1}{8}(17 - (65)^{1/2}) = 1.117\ldots. \tag{3.6}$$

This last estimate is the sharpest known at present. A simplified and concise[1] version of the proof of (3.5) is given in [5].

4. Proofs of (3.1)

The proof of Erdös, mentioned in Section 3, was as follows.

We may take n to be a prime p. (This involves no loss of generality, since the ratio of consecutive primes is asymptotically 1.) We take the region $\mathbf{K}$ to be the square $0 \leqslant x \leqslant p$, $0 \leqslant y \leqslant p$ in the (x, y) plane. For each $\nu = 1, 2,\ldots, p$, we define a point $P_\nu = (x_\nu, y_\nu)$ in $\mathbf{K}$, with integer coordinates, such that

$$x_\nu \equiv \nu, \qquad y_\nu \equiv \nu^2 \qquad (\mathrm{mod}\, p).$$

Now, if $1 \leqslant i < j < k \leqslant p$, the determinant

$$\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix} \equiv \begin{vmatrix} 1 & i & i^2 \\ 1 & j & j^2 \\ 1 & k & k^2 \end{vmatrix} \qquad (\mathrm{mod}\, p)$$

cannot vanish, so that $\triangle^*(\mathbf{K}; p) \geqslant \frac{1}{2}$ and hence $\triangle(\mathbf{S}, p) \geqslant \frac{1}{2}p^{-2}$.

[1] [5] embodies a number of genuine simplifications of the proof. But on the few occasions when phrases such as "it is easily seen" are used in [5], the appropriate details are to be found in [3] or [4].

Although the above construction is simple and elegant, it is more natural and less restrictive to select the points $P_1, P_2, \ldots, P_n$ one at a time, at each stage ensuring that (for a suitably large positive constant c)

(i) *at each insertion of a "new" P_ν, no triangle (of type $P_iP_jP_k$) of area less than $c^{-1}n^{-2}$ is formed,*

(ii) *no two points P_i shall be too close together (so that no pair P_i, P_j "prohibits" too much space for subsequent selections of P_ν).*

We shall describe this procedure more fully, and for this purpose take $\mathbf{K}$ to be a disc $\mathbf{D}$ of area 1. We may take the requirement (ii) to be (using $| P - Q |$ to denote the distance between P and Q)

$$| P_i - P_j | > \tfrac{1}{4}n^{-1/2} \qquad (i < j). \tag{4.1}$$

At the νth stage of the construction, when P_ν is selected, the space (within $\mathbf{D}$) "prohibited" (in view of (i) and (4.1)) by

$$P_1, P_2, \ldots, P_{\nu-1} \tag{4.2}$$

consists of the union of

(I) *the open strip of width $4c^{-1}n^{-2} | P_i - P_j |^{-1}$ about the (central) line P_iP_j (joining P_i, P_j) corresponding to each pair P_i, P_j with $1 \leqslant i < j \leqslant \nu - 1$,*

(II) *the disc of radius $\frac{1}{4}n^{-1/2}$ about each point P_i with $1 \leqslant i \leqslant \nu - 1$.*

The space (within $\mathbf{D}$) "prohibited" in this way has total area less than

$$\nu(\tfrac{1}{4}n^{-1}) + 8c^{-1}n^{-2} \sum_{1 \leqslant i < j \leqslant \nu-1} | P_i - P_j |^{-1}. \tag{4.3}$$

When P_ν is to be selected, (4.1) is already available for $1 \leqslant i < j \leqslant \nu - 1$. Using this (and the fact that c is large) to estimate the sum appearing in (4.3), it is easily verified that provided $\nu < n$, the value of the expression (4.3) is less than 1 (the area of $\mathbf{D}$). Thus the construction cannot break down before it is complete.

The above proof was included in Schmidt's paper [2]. But the argument may well have been discovered independently by most of those mathematicians who have made unsuccessful attempts to disprove Heilbronn's conjecture. If one could impose further restrictions[2] on the P_ν so as to reduce the area of the *union* of the strips (I), a disproof of Heilbronn's conjecture might result.

[2] Possibly in conjunction with some averaging argument.

5. Notation

We introduce some further notation, to be used *in addition to that introduced in Section* 2; in particular, we recall from the final paragraph of Section 2 that *n is supposed to be large throughout* and that $\triangle(n)$ stands for $\triangle(\mathbf{D}; n)$.

We use $X = (x, y)$ to denote a point in the Euclidean plane, and write $\int g(X)\, dX$ for $\iint g(x, y)\, dx\, dy$ taken over the entire plane.

If $\mathbf{V}$ is a subset of the plane, we use $\mathbf{V}(X)$ to denote the characteristic function of $\mathbf{V}$; in other words $\mathbf{V}(X)$ is 1 or 0 according as X does or does not lie in $\mathbf{V}$. (Thus, for example, the area $A(\mathbf{K})$ of the convex region $\mathbf{K}$ is given by $A(\mathbf{K}) = \int \mathbf{K}(X)\, dX$).

We use λX, $\mu \mathbf{V}$ (where λ is real and $\mu > 0$) in the usual way; thus $\lambda X = (\lambda x, \lambda y)$ and $\mu \mathbf{V}$ denotes the set with characteristic function $\mathbf{V}(\mu^{-1} X)$. We use $| X |$ to denote the distance of X from the origin, so that $| X' - X'' |$ is the distance between X' and X''.

We intruduce some *notation relative to the extremal set* (2.1) *appearing in Section* 2.

We use τ to denote a pair P_i, P_j $(i < j)$ of points selected from the set (2.1), and $d(\tau)$ to denote the distance $| P_i - P_j |$ between the two points constituting τ. If

$$y \cos \theta - x \sin \theta = a \qquad (0 \leqslant \theta < \pi) \tag{5.1}$$

is the line joining the constituent points P_i, P_j of τ, we use $\theta(\tau)$ to denote the inclination θ of the line (5.1). Furthermore, for any $w > 0$, we use $\mathbf{H}_\tau(w)$ to denote the (open) strip

$$a - \tfrac{1}{2}w < y \cos \theta - x \sin \theta < a + \tfrac{1}{2}w \tag{5.2}$$

of width w about the line (5.1).

We extend the above notation for characteristic functions, by writing $\mathbf{H}_\tau(w; X)$ for the characteristic function of $\mathbf{H}_\tau(w)$. We denote by $N_\tau(w)$ the number of points of the set (2.1) lying in $\mathbf{H}_\tau(w)$, so that

$$N_\tau(w) = \sum_{i=1}^{n} \mathbf{H}_\tau(w; P_i). \tag{5.3}$$

Statements concerning "every pair τ" refer to the set of $\tfrac{1}{2}n(n - 1)$ possible selections of such pairs, and $\sum_\tau$ denotes a summation over this set.

We use c as a generic symbol for a *sufficiently large* absolute constant.

6. Properties of the Extremal Distribution (2.1) if $n\triangle(\mathbf{K}; n)$ Is Not "Small"

In this section we suppose that $\mathbf{K}$ *is a disc, or a square, or an equilateral triangle.* We suppose further that

$$A(\mathbf{K}) = 1 \tag{6.1}$$

and that the centroid of $\mathbf{K}$ is at the origin (so that $\mathbf{K}$ and $2\mathbf{K}$ are concentric). We shall write simply $\triangle$ for $\triangle(\mathbf{K}; n)$, so that (in this section only) $\triangle$ stands for $\triangle(\mathbf{D}; n)$, $\triangle(\mathbf{S}; n)$ or $\triangle(\mathbf{T}; n)$ according as $\mathbf{K}$ is a disc, a square, or a triangle.

Heilbronn's problem is that of showing that $n\triangle$ is "small." We now discuss the consequences upon the extremal distribution (2.1) of the contrary assumption that $n\triangle$ is not "small" (in some appropriate sense); our purpose being to exhibit properties that might be useful in deducing a contradiction.

Much of the discussion in this section will be heuristic or purely descriptive in nature, and is not intended to be precise in detail. In particular, we largely ignore complications that can arise in relation to pairs τ lying close to the boundary of $\mathbf{K}$, and various assertions relating to pairs τ are in fact justified only if such exceptional τ are excluded.

The fact that no three of the points (2.1) form a triangle of area less than $\triangle$ is equivalent to saying that, for each pair τ, the strip

$$\mathbf{H}_\tau(4\triangle/d(\tau)) \tag{6.2}$$

contains no member of the set (2.1) other than the two points P_i, P_j constituting τ; in other words

$$N_\tau(4\triangle/d(\tau)) = 2 \qquad \textit{for every } \tau. \tag{6.3}$$

Since the distribution (2.1) consists of n points in a region $\mathbf{K}$ of area 1, the (statistical) expectation for $N_\tau(4\triangle/d(\tau))$ is

$$n \int \mathbf{K}(X)\,\mathbf{H}_\tau(4\triangle/d(\tau); X)\,dX; \tag{6.4}$$

that is n times the area of that part of $\mathbf{K}$ lying in the strip. This area is in general between constant factors of $\Delta/d(\tau)$, so that we may take the expectation to have roughly the order of magnitude

$$n\triangle/d(\tau). \tag{6.5}$$

Now if $n\triangle$ is not "small" (for example if $n\triangle > n^{-0.001}$) but $d(\tau)$ is very small (for example if $d(\tau) < cn^{-1/2}$) the expectation for $N_\tau(4\triangle/d(\tau))$ is large, whilst the actual value 2 given by (6.3) is bounded. We may describe this phenomenon by saying that, if $d(\tau)$ is very small, the strip (6.2) is very "deficient" of points P_i of the distribution (2.1). In other words, when $d(\tau)$ is very small the strip (6.2) takes up an undue proportion of the area of $\mathbf{K}$ in relation to its content of points P_i. This circumstance is clearly highly relevant to our purpose; for less space is available to the remaining P_i. If, for example, there is a convex region $\mathbf{K}_1$ containing substantially more (by a large factor) than its "expectation" of points P_i, we could deduce the existence of a triangle of small area (with vertices among (2.1)) in $\mathbf{K}_1$.

We can sum up the above discussion by noting that, on the assumption that $n\triangle$ is not "small," the distribution (2.1) has a property of the following general nature.

PROPERTY (A). *The strips $\mathbf{H}_\tau(4\triangle/d(\tau))$ corresponding to pairs τ the constituent points of which are "close together" (i.e., with very small $d(\tau)$) are very "deficient" of points P_i (and, in this sense, each such strip takes up an undue amount of space in $\mathbf{K}$).*

It is clear from the above discussion that the manner in which the strips $\mathbf{H}_\tau$, corresponding to small $d(\tau)$, overlap is relevant to our problem. But here again, we can easily utilize our assumption that $n\triangle$ is not "small." Each strip $\mathbf{H}_\tau$ corresponds to a pair τ, and no two τ_1, τ_2 of these pairs are situated so that a triangle of area less than $\triangle$ is formed by any three of the four constituent points of τ_1, τ_2. It is easy to establish the following concrete consequence of this remark.

LEMMA B_1. *Let $u > 0$, and suppose that the pairs τ_1, τ_2 are such that*

$$d(\tau_1) \leqslant u, \qquad d(\tau_2) \leqslant u, \qquad |\,\theta(\tau_1) - \theta(\tau_2)| < 10^{-2}\triangle u^{-1}.$$

Then the strips

$$\mathbf{H}_{\tau_1}(10^{-2}\triangle u^{-1}), \qquad \mathbf{H}_{\tau_2}(10^{-2}\triangle u^{-1})$$

do not overlap within $\mathbf{K}$.

It can be deduced from Lemma B_1 that, on the assumption that $n\triangle$ is not small, suitable systems of strips of type $\mathbf{H}_\tau(w)$ (with any fixed w exceeding an appropriate bound) have the property that the number of times a point X_1 of $\mathbf{K}$ is covered by strips of a given system $\mathscr{S}$ does not exceed by too great a factor the corresponding statistical "expectation," namely,

$$\sum_{\tau;\mathbf{H}_\tau\in\mathscr{S}} \int \mathbf{H}_\tau(w;\,X)\,\mathbf{K}(X)\,dX. \tag{6.6}$$

To avoid complications that can arise in connection with pairs τ close to the boundary of $\mathbf{K}$, it is usually technically more convenient to work with expressions of type

$$\sum_{\tau;\mathbf{H}_\tau\in\mathscr{S}} \int \mathbf{H}_\tau(w;\,X)\,\mathbf{K}(\tfrac{1}{2}X)\,dX \tag{6.7}$$

in place of (6.6). Accordingly, when estimating the number of times a typical point X_1 is covered by strips of $\mathscr{S}$, it is preferable to take into consideration all X_1 in the enlarged region $2\mathbf{K}$.

After Lemma B_1 has been slightly modified (by replacing $\mathbf{K}$ by $2\mathbf{K}$ in the final sentence) the following is representative of results easily deduced from it.

LEMMA B_2. *Let $u > 0$, $0 \leqslant \alpha < \pi$, $w \geqslant \frac{1}{2}\triangle u^{-1}$. Then, for a suitable constant c,*

$$\mathbf{K}(\tfrac{1}{2}X) \sum_{\tau;(6.9)} \mathbf{H}_\tau(w;\,X) \leqslant cwu\triangle^{-1} \tag{6.8}$$

for every X in the plane: here the summation is over all pairs τ satisfying

$$d(\tau) \leqslant u, \qquad \alpha \leqslant \theta(\tau) < \alpha + 10^{-2}\triangle u^{-1}. \tag{6.9}$$

This lemma provides a typical example of the type of result we have described. When X is in $2\mathbf{K}$, the left-hand side of (6.8) represents the number $k(\mathscr{S};\,X)$ of times the point X is covered by strips of the system $\mathscr{S}$ consisting of all $\mathbf{H}_\tau(w)$ with τ satisfying (6.9). The expectation for this number has roughly the order of magnitude $wun^2\triangle$, and the right-hand side of (6.8) does not exceed this by a large factor unless $n\triangle$ is small.

Again summing up in descriptive language, we have noted that if $n\triangle$ is not "small," the strips of type $\mathbf{H}_\tau$ have a property of the following general nature.

PROPERTY (B). *Suppose that $\mathscr{S}$ is a suitable system of strips of type $\mathbf{H}_\tau(w)$. Then the number $k(\mathscr{S};\,X)$ of times an arbitrary point X of $2\mathbf{K}$ is* *covered by strips of $\mathscr{S}$ will not exceed by too large a factor the statistical expectation for this number.*

(Here the word "suitable" signifies that $\mathscr{S}$ is assumed to satisfy appropriate conditions. Such condition must ensure that w is not too small, but need not be unduly restrictive in other respects.)

All known methods for obtaining upper bounds for $\triangle$ are based (explicitly or implicitly) on the exploitation of properties of type (A) and (B) enjoyed by extremal distributions (2.1) when $n\triangle$ is not too small.

7. Roth's Early Method

We now describe, very briefly, the central ideas of Roth's proof of (3.3); in doing so we suppress (among others) some tedious minor complications that arise in relation to pairs τ lying close to the boundary of $\mathbf{K}$.

$\mathbf{K}$ is taken to be an equilateral triangle $\mathbf{T}$ of area 1. There are at least $c_1^{-1}n$ pairs τ satisfying

$$d(\tau) \leqslant 10^{-1}n^{-1/2} \tag{7.1}$$

and (by a box argument) for some α there are at least $c_2^{-1}n^{3/2}\triangle(\mathbf{T}; n)$ of these which also satisfy

$$\alpha \leqslant \theta(\tau) < \alpha + 10^{-1}n^{1/2}\triangle(\mathbf{T}; n). \tag{7.2}$$

By Lemma B_1 of Section 6 (with $u = 10^{-1}n^{-1/2}$), the set $\mathscr{S}$ of strips

$$\mathbf{H}_\tau(10^{-1}n^{1/2}\triangle(\mathbf{T}; n))$$

satisfying both (7.1) and (7.2) intersect $\mathbf{K}$ in disjoint regions; and, as was explained in Section 6 (cf. Property (A)), these regions are "deficient" of points P_i (unless $n\triangle(\mathbf{T}; n)$ is "small," as desired). The part of the triangle $\mathbf{T}$ that remains after the above-mentioned regions are excluded consists of (at[3] most $|\mathscr{S}| + 1$) convex polygons (each having at most six sides), and these contain between them more than the "expected" number of points P_i (again, unless $n\triangle(\mathbf{T}; n)$ is "small"). On decomposing these convex polygons into triangles (which yields a total of at most $4(|\mathscr{S}| + 1)$ triangles), we see there is a triangle $\mathbf{T}_1$ containing "too

[3] $|\mathscr{S}|$ denotes the number of elements of the set $|\mathscr{S}|$.

many" points P_i in relation to its area. This phenomenon may be expressed in terms of a functional inequality for $\triangle(\mathbf{T}; n)$, by relating $\triangle(\mathbf{T}; n)$ to $\triangle(\mathbf{T}; n_1)$, where n_1 is the number of points P_i in $\mathbf{T}_1$. The estimate (3.3) is a consequence of such a functional inequality.

This method is too weak to remain of interest now, except as a possible source of ideas which might be applicable in conjunction with more powerful procedures. The main point of interest is the transition from the equilateral triangle $\mathbf{T}$ to the triangle $\mathbf{T}_1$ which must (implicitly) again be transformed, by a linear transformation of the plane, into an equilateral triangle. Of the known procedures for obtaining upper bounds for $\triangle(n)$, the above method is the only one that makes use of general[4] linear transformations of the plane. If the use of such transformations could in some way be combined with more recent methods, new results might ensue.

8. The Method of Schmidt

In the following account $\mathbf{K}$ may, following Schmidt, be taken to be a square of area 1; but it can equally well be taken to be a disc of area 1. As in Section 6, we suppose that the center of $\mathbf{K}$ is at the origin (so that $\mathbf{K}$ and $2\mathbf{K}$ are concentric) and we shall write simply $\triangle$ for $\triangle(\mathbf{K}; n)$.

It was remarked in Section 6 (see Property (B)) that if $n\triangle$ is not "small," then appropriate systems of strips of type $\mathbf{H}_\tau(w)$ do not "cover" any point X of $2\mathbf{K}$ an unduly large number of times. Schmidt's method is based on a more sophisticated observation of the same general nature.

Let c_0 be a suitably large constant. To each τ associate a strip

$$\mathbf{H}_\tau(w_\tau), \qquad \textit{where} \quad w_\tau = \triangle/(c_0 d(\tau)). \tag{8.1}$$

We shall suppose that the w_τ are small (less than c^{-1}), an assumption that is easily justified for the purpose of the method.

The essence of Schmidt's idea was to apply the following result in the context of weighted means.

Lemma. *If the strips* (8.1) *corresponding to s pairs $\tau_1, \tau_2, \dots, \tau_s$ have a point X_1 of* $2\mathbf{K}$ *in common, then*

$$w_{\tau_1} + w_{\tau_2} + \cdots + w_{\tau_s} < \pi. \tag{8.2}$$

[4] The fact that $\triangle(\mathbf{D}; n)$ is independent of the disc $\mathbf{D}$ has been used in a later method (cf. Sect. 10, (10.6)), but the linear transformation implicit in this invariance (a mere "magnification") is of a very special kind.

This lemma is proved as follows. Let C be the circle center X_1 and radius 1. Since the w_τ are small, the strip $\mathbf{H}_{\tau_j}(w_{\tau_j})$ intersects C in two arcs each of length greater than w_{τ_j}. Thus, assuming that (8.2) is false, there is a point X_2 on C which lies in two of the s strips under consideration. In other words two of the strips, which we may take to be $\mathbf{H}_{\tau_1}(w_{\tau_1})$ and $\mathbf{H}_{\tau_2}(w_{\tau_2})$, have both the points X_1 and X_2 in common. Supposing (as we may) that $w_{\tau_1} \geqslant w_{\tau_2}$, it follows that the widened strip $\mathbf{H}_{\tau_1}(c_0 w_{\tau_1})$ contains the constituent points of the pair τ_2 (since c_0 is suitably large); and this is a contradiction, since the two points of τ_1 together with one of the points of τ_2 now define a triangle of area less than Δ.

The assertion of the lemma may be expressed as follows.

$$\mathbf{K}(\tfrac{1}{2}X) \sum_\tau w_\tau \mathbf{H}_\tau(w_\tau\,;\, X) < \pi \tag{8.3}$$

for all X in the plane; here, of course, the summation is over *all* pairs τ.

On the other hand it is easily shown that, if $L(X)$ denotes the left-hand side of (8.3), then

$$\int L(X)\, dX \gg \sum_\tau w_\tau^{\,2} \gg \Delta^2 \sum_\tau \frac{1}{(d(\tau))^2} \gg \Delta^2 n^2 \log n.$$

Thus on integrating (8.3) we at once obtain (3.4), namely, $\Delta \ll n^{-1}(\log n)^{-1/2}$.

The later method of Roth is again based on the use of "weighted" strips, although, as we shall see, both the nature and purpose of the "weighting" is entirely different.

9. Roth's Recent Method: The Main Idea

In this section $\mathbf{K}$ is taken to be[5] a disc $\mathbf{D}$ of area 1. We again suppose that the center of $\mathbf{D}$ is at the origin (so that $\mathbf{D}$ and $2\mathbf{D}$ are concentric) and shall write simply Δ in place of $\Delta(\mathbf{D}; n)$.

Roth's recent method made use, for the first time, of the special nature of the region of intersection of any two strips $\mathbf{H}_{\tau^*}(w)$, $\mathbf{H}_{\tau^{**}}(W)$. The fact that (for given τ^*, τ^{**}) the area

$$wW \,|\operatorname{cosec}(\theta(\tau^*) - \theta(\tau^{**}))| \tag{9.1}$$

[5] There is a conflict between the notation adopted here and that used in [5]. In [5], $\mathbf{D}^*$ denoted a disc of area 1 and $\mathbf{D}$ was used to denote a certain larger disc; but here $\mathbf{D}$ denotes a disc of area 1 and a larger disc (namely, $2\mathbf{D}$) will be introduced implicitly via the use of the notation $\mathbf{D}(\tfrac{1}{2}X)$.

of the parallelogram of intersection (if finite) is proportional to each of w and W enables one to construct systems of orthogonal functions from "weighted" strips; and we shall see that an application of Bessel's inequality to (a modified form of) such orthogonal systems enables one to exploit much more effectively the properties (A) and (B) described in Section 6.

Suppose that, using the informal terminology of Section 6, $n\triangle$ is not "small." (The actual requirement here is of the type $n\triangle > n^{-1/c}$ for a suitable constant c.) Then, according to Property (A), pairs τ with small $d(\tau)$ generate certain strips "deficient" of points P_i. Consider, for example, the set of those τ satisfying

$$d(\tau) \leqslant n^{-1/3}. \tag{9.2}$$

It follows from (6.3) that the strips

$$\mathbf{H}_\tau(n^{1/3}\triangle) \tag{9.3}$$

corresponding to such τ, each contain exactly two points P_i; whereas, in the sense of Section 6, the corresponding expectation for the number of points P_i has (in general) the order of magnitude

$$n^{1/3}(n\triangle) \tag{9.4}$$

and is therefore large.

In general, the expectation for the number of points P_i in a strip $\mathbf{H}_\tau(w)$ will have the order of magnitude nw, so that (cf. (5.3))

$$n^{-1}w^{-1}N_\tau(w) \tag{9.5}$$

is the order of magnitude of the ratio of the actual number to the expected number of such points.

We have remarked above that when $w = n^{1/3}\triangle$ (which would be less than $n^{-2/3}$), the ratio (9.5) is very small for all τ. No direct procedure is known for obtaining a contradiction from this. The reason is that when w is as small as $n^{-2/3}$ there is no way of selecting pairs τ for which the ratio (9.5) is not small. But when w is larger, say $w = n^{-1/4}$, such procedures are feasible, and indeed it can be shown that in this case the ratio (9.5) exceeds a positive constant for a significant proportion of the set of all pairs τ satisfying (9.2). We shall not discuss the technicalities of these procedures (or even state in detail the results that ensue) and confine ourselves to a purely descriptive account of the underlying principle on which they are based.

Suppose that $\mathbf{S}$ is any strip[6] of type[7]

$$-\tfrac{1}{2}c^{-1}w < x\cos\alpha - y\sin\alpha - a \leqslant \tfrac{1}{2}c^{-1}w$$

and suppose further that it is known that $\mathbf{S}$ contains roughly the "expected" number $c^{-1}nw$ of points P_i. Then, if a strip $\mathbf{H}_\tau(w)$ covers the set $\mathbf{D} \cap \mathbf{S}$ (in which $\mathbf{S}$ intersects the disc $\mathbf{D}$) it cannot be unduly deficient of points P_i; for $\mathbf{H}_\tau(w)$ will certainly contain all the points P_i of $\mathbf{S}$. The condition that $\mathbf{H}_\tau(w)$ should cover the set $\mathbf{D} \cap \mathbf{S}$ will certainly be satisfied if (for suitably large c) both

$$\tau \text{ lies in } \mathbf{S} \text{ (i.e., both points of } \tau \text{ lie in } \mathbf{S}), \tag{9.6}$$

$$|\,\theta(\tau) - \alpha\,| < c^{-1}w. \tag{9.7}$$

Strips $\mathbf{S}$ containing roughly the expected number of points P_i are easily constructed by subdividing the plane (or possibly a previously constructed strip whose width is a multiple of w and which is not deficient of points P_i) into strips of width w (and the same inclination α) and then rejecting those "bad" strips in the subdivision which are unduly deficient of points P_i. Providing that w is not too small, it is possible to show[8] that there is a numerous supply of pairs τ (of the kind[9] under consideration) with each τ lying in one of the remaining "good" strips $\mathbf{S}$ and also satisfying (9.7) above.

On averaging over α we can obtain in this way, for suitable w', results of the following general nature.

(I) *A significant proportion of all those τ for which $d(\tau)$ is appropriately restricted are such that the strips $\mathbf{H}_\tau(w')$ are not unduly deficient of points P_i.*

On the other hand, the procedure described earlier yields, for suitable u and w'', a result of the following type.

(II) *All the strips $\mathbf{H}_\tau(w'')$ for which*

$$d(\tau) \leqslant u \tag{9.8}$$

are very deficient of points P_i.

[6] This notation should not cause any confusion, as there will be no further reference to squares.

[7] Here the central line of $\mathbf{S}$ is not generated by a pair τ, so that the interior of $\mathbf{S}$ need not be of type $\mathbf{H}_\tau(c^{-1}w)$.

[8] For this purpose also, some use is made of the assumption that $n\triangle$ is not "small."

[9] Say, for example, satisfying (9.2).

The two results are available only subject to appropriate premises, and these are never satisfied when (I) and (II) would be directly contradictory. If the restriction on $d(\tau)$ in (I) is taken to be identical to that in (II) (namely, $d(\tau) \leqslant u$) the respective premises require that the order of magnitude of w' is large compared to that of w''; a typical admissible choice of parameters in this case would be $u = n^{-1/3}$, $w' = n^{-1/4}$, $w'' = n^{-3/4}$ (if $n\triangle > n^{-1/12}$).

The hard core of Roth's method, in its simplest form[10], consists of a technique for deducing from (II) a result that contradicts (I) (again on the assumption that $n\triangle$ is not small). In other words (after a suitable admissible choice of parameters) it is deduced from the fact that all the strips $\mathbf{H}_\tau(w'')$ (with τ subject to (9.8)) are very deficient of points P_i, that almost all the wider strips $\mathbf{H}_\tau(w')$ (again with τ subject to (9.8)) are deficient of points P_i; this latter "deficiency" may be less severe than that of the $\mathbf{H}_\tau(w'')$, but still suffices to contradict (I). The deduction is based on the application of Bessel's inequality, to systems of (modified) orthogonal functions, already mentioned at the beginning of the section.

Suppose that $w' > w'' > 0$. Consider two functions of type

$$\frac{1}{w}\mathbf{H}_{\tau^*}(w; X), \qquad \frac{1}{W}\mathbf{H}_{\tau^{**}}(W; X); \tag{9.9}$$

we may interpret these as the characteristic functions of the strips $\mathbf{H}_{\tau^*}(w)$, $\mathbf{H}_{\tau^{**}}(W)$ "weighted" in inverse proportion to their widths. The total weight associated to the product of the two function (9.9), namely,

$$\int \frac{1}{wW}\mathbf{H}_{\tau^*}(w; X)\,\mathbf{H}_{\tau^{**}}(W; X)\,dX \tag{9.10}$$

is independent of w, W (cf. the earlier remark concerning (9.1)). Thus the functions

$$\phi_\tau(w', w''; X) = \frac{1}{w'}\mathbf{H}_\tau(w'; X) - \frac{1}{w''}\mathbf{H}_\tau(w''; X) \tag{9.11}$$

have the following property.

ORTHOGONALITY PROPERTY. *If ϕ^*, ϕ^{**} are any two functions of type* (9.11) (*corresponding to pairs of τ^*, τ^{**}*), *then $\int \phi^*\phi^{**}\,dX$ is zero whenever it is finite.*

[10] That is devoid of further devices superimposed to effect a quantitative improvement of the final result.

As we have in mind an application of Bessel's inequality, we must work with functions of finite norms, and indeed will require these norms to be not too large. Accordingly, we replace the functions (9.11) by the modified functions

$$\Phi_\tau(w', w''; X) = \mathbf{D}(\tfrac{1}{2}X)\,\phi_\tau(w', w''; X), \tag{9.12}$$

derived by replacing ϕ_τ by zero outside the disc $2\mathbf{D}$. (As will become clear in due course, the use of the larger disc $2\mathbf{D}$ in preference to $\mathbf{D}$ is again designed to avoid complications arising in connection with points P_i near the boundary of $\mathbf{D}$.)

The modified functions (9.12) no longer form a fully orthogonal system, but the system is quasi-orthogonal in the following sense.

QUASI-ORTHOGONALITY PROPERTY. *For given $w' > w'' > 0$, two functions Φ_{τ^*}, $\Phi_{\tau^{**}}$ are orthogonal over the plane unless the common parallelogram of the strips $\mathbf{H}_{\tau^*}(w')$, $\mathbf{H}_{\tau^{**}}(w')$ intersects the boundary circle of the disc $2\mathbf{D}$.*

This quasi-orthogonality property can be made to suffice for our purpose because, if τ^* is fixed and τ^{**} runs through the set of τ under consideration, the exceptional case in which Φ_{τ^*}, $\Phi_{\tau^{**}}$ are not orthogonal occurs only comparatively rarely and with the values of $\theta(\tau^{**})$ not too "bunched" (cf. Sect. 6, Lemma B_2). We will, however, require an appropriately modified form of Bessel's inequality applicable to quasi-orthogonal systems. Such generalizations, which have proved very fruitful in connection with the large sieve, are discussed in [6, Chap. 1]. The generalized Bessel's inequality actually used is due to A. Selberg (see [6, p. 7]); the proof [6, p. 8] of Selberg's elegant result being at least as simple as that of weaker inequalities of the same general nature (some of which would suffice for our purpose).

SELBERG'S INEQUALITY. *Let f, $\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(R)}$ be elements of an inner product space over the complex numbers. Then*

$$\sum_{r=1}^{R} |(f, \psi^{(r)})|^2 \left(\sum_{s=1}^{R} |(\psi^{(r)}, \psi^{(s)})| \right)^{-1} \leqslant \|f\|^2. \tag{9.13}$$

In our application, we will of course be concerned with inner products of type

$$(f, g) = \int f(X)\, g(X)\, dX,$$

where f, g are real functions over the plane; and, as always, $\|f\|^2 = (f,f)$.

For each $i = 1, 2, \ldots, n$, let g_i denote the function which is 1 or 0 according as X does or does not lie in the *disc* $\mathbf{D}^{(P_i)}$ *center* P_i *and radius* $\frac{1}{2}w''$. We may express this symbolically by

$$\mathbf{D}^{(P_i)} = P_i + (\tfrac{1}{2}\pi^{1/2}w'')\mathbf{D}, \qquad g_i(X) = \mathbf{D}^{(P_i)}(X). \tag{9.14}$$

For each i the disc $\mathbf{D}^{(P_i)}$ has area $A = \pi(\frac{1}{2}w'')^2$, so that

$$A = \pi(\tfrac{1}{2}w'')^2 = \int g_i(X)\, dX \qquad (i = 1,\ldots, n). \tag{9.15}$$

We write

$$f(X) = \sum_{i=1}^{n} g_i(X). \tag{9.16}$$

We may think of $f(X)$ as being a mass distribution which approximates that obtained by placing a mass A at each of the points P_i; each mass has been spread over a disc in order to reduce the norm of f. (It would not be practical to spread each mass over a disc of radius significantly larger than $\frac{1}{2}w''$, because we want the total mass falling into a typical strip of width w'' to be roughly proportional to the number of points P_i in the strip.)

If we ignore for the moment the error due to the fact that the discs $\mathbf{D}^{(P_i)}$ may overlap the edges of the relevant strips (i.e., if we treat the g_i as though they correspond to point masses), we surmise that for each τ,

$$\int \Phi_\tau(w', w''; X) f(X)\, dX \tag{9.17}$$

is approximately (cf. (5.3), (9.11), (9.12), (9.15))

$$\pi(\tfrac{1}{2}w'')^2 \left\{\frac{1}{w'} N_\tau(w') - \frac{1}{w''} N_\tau(w'')\right\}. \tag{9.18}$$

With the abbreviated notation $\Phi_\tau = \Phi_\tau(w', w''; X), f = f(X)$, we now apply Selberg's inequality (9.13) in the form

$$\sum_{\tau^*; d(\tau^*)\leqslant u} \left| \int \Phi_{\tau^*} f\, dX \right|^2 \left(\sum_{\tau^{**}; d(\tau^{**})\leqslant u} \left| \int \Phi_{\tau^*}\Phi_{\tau^{**}}\, dX \right|\right)^{-1} \leqslant \int f^2\, dX. \tag{9.19}$$

Lemma B_2 (see Sect. 6) can be used to estimate (for any fixed τ^*) the sum

$$\sum_{\tau^{**}; d(\tau^{**}) \leqslant u} \left| \int \Phi_{\tau^*}(X)\, \Phi_{\tau^{**}}(X)\, dX \right| \tag{9.20}$$

appearing on the left-hand side of (9.19). Of course, only the "exceptional" τ^{**} for which the orthogonality breaks down contribute to the sum (9.20); these are the τ^{**} for which the strip $\mathbf{H}_{\tau^{**}}(w')$ intersects one of the two small arcs which the (fixed) strip $\mathbf{H}_{\tau^*}(w')$ has in common with the boundary circle of the disc $2\mathbf{D}$. The bound for the sum (9.20) which results from the application of Lemma B_2 is a good one if $n\triangle$ is not small. The estimation of $\int f^2\, dX$ does not give rise to any difficulty. Thus, if we were entitled to replace (9.17) by (9.18), the application of the modified Bessel's inequality (9.19) would yield a good upper bound for

$$\sum_{\tau; d(\tau) \leqslant u} \left\{ \frac{1}{w'} N_\tau(w') - \frac{1}{w''} N_\tau(w'') \right\}^2. \tag{9.21}$$

By "good" we here mean an upper bound which is small compared to the "expectation" of roughly

$$\sum_{\tau; d(\tau) \leqslant u} n^2 \qquad \text{for the sum} \qquad \sum_{\tau; d(\tau) \leqslant u} \left\{ \frac{1}{w'} N_\tau(w') \right\}^2.$$

Such an estimate for the sum (9.21) would enable us to deduce from the fact that $(1/w'')\, N_\tau(w'')$ is always small compared to its expected value, that $(1/w')\, N_\tau(w')$ is nearly always small compared to its expected value[11] (where in each case τ is subject to $d(\tau) \leqslant u$). This is the desired deduction, from (II), of a result that contradicts (I).

The above argument fails to deal with the difficulty of justifying the transition from (9.17) to (9.18), and in fact it is not practicable to estimate satisfactorily the error term this transition would introduce. The difficulty can however be avoided altogether by the following slight modification of the argument.

Suppose now that

$$0 < w'' < \tfrac{1}{4} w' < c^{-1}; \tag{9.22}$$

[11] Each of these expected values has roughly the order of magnitude n.

these inequalities are amply satisfied in the application we have in mind. We recall that the disc $\mathbf{D}^{(P_i)}$, in which the function g_i assumes the value 1, has center P_i and radius $\frac{1}{2}w''$. This disc will lie entirely within the strip $\mathbf{H}_\tau(w')$ provided its center P_i lies in $\mathbf{H}_\tau(w' - w'')$, and in particular if P_i lies in $\mathbf{H}_\tau(\frac{1}{2}w')$. Hence, by (9.15) and (5.3),

$$\int \mathbf{H}_\tau(w'; X) f(X)\, dX \geqslant AN_\tau(\tfrac{1}{2}w'). \tag{9.23}$$

On the other hand the disc $\mathbf{D}^{(P_i)}$ will intersect the strip $\mathbf{H}_\tau(w'')$ only if P_i lies in $\mathbf{H}_\tau(2w'')$, and thus

$$\int \mathbf{H}_\tau(w''; X) f(X)\, dX \leqslant AN_\tau(2w''). \tag{9.24}$$

Combining (9.23), (9.24), and noting that the set in which $f(X)$ is nonzero is contained in the disc $2\mathbf{D}$ in which $\mathbf{D}(\frac{1}{2}X) = 1$, we obtain (on recalling the definition (9.11), (9.12) of Φ_τ),

$$\int \Phi_\tau(w', w''; X) f(X)\, dX \geqslant A \left\{ \frac{1}{w'} N_\tau(\tfrac{1}{2}w') - \frac{1}{w''} N_\tau(2w'') \right\}. \tag{9.25}$$

On applying (9.19) exactly as before, we now obtain a bound for the sum

$$\sum_{\tau; d(\tau) \leqslant u}^{*} \left\{ \frac{1}{w'} N_\tau(\tfrac{1}{2}w') - \frac{1}{w''} N_\tau(2w'') \right\}^2,$$

where the asterisk signifies that the summation is restricted to those τ for which

$$\frac{1}{w'} N_\tau(\tfrac{1}{2}w') - \frac{1}{w''} N_\tau(2w'') > 0.$$

On writing $W' = \frac{1}{2}w'$, $W'' = 2w''$, we see that subject to $0 < W'' < W' < c^{-1}$, we have obtained an estimate for

$$\sum_{\tau; d(\tau) \leqslant u}^{**} \left\{ \frac{1}{W'} N_\tau(W') - \frac{4}{W''} N_\tau(W'') \right\}^2, \tag{9.26}$$

where the double asterisk signifies that the summation is restricted to those τ for which the bracketed expression in the summand is positive.

For the application we have described, we require that $(1/w')\, N_\tau(w')$ cannot too often be "large" whilst $(1/w'')\, N_\tau(w'')$ is "small." For this purpose a good bound for the sum (9.26) (after a change of notation consisting of the replacement of W', W'' by w', w'') is entirely adequate. There is therefore no need to estimate sums of type (9.21), and the transition from (9.17) to (9.18) (which provided the heuristic motivation for our main strategy) can be avoided altogether.

10. Roth's Recent Method: Further Remarks

In Section 9 we described Roth's recent method in its simplest form. We now give some indication of the devices that are superimposed in order to improve the resulting estimate for $\triangle$.

To deduce (from (II)) a result in contradiction to (I), we wish to estimate for $0 < \delta < 1$, the number of τ for which

$$d(\tau) \leqslant u, \qquad N_\tau(w') \geqslant \delta n w'; \tag{10.1}$$

let us denote by $B(u;\, \delta,\, w')$ the number of such τ.

We have at our disposal (see (9.26)) an estimate of type

$$\sum_{\tau;\, d(\tau) \leqslant u}^{**} \left\{ \frac{1}{W'}\, N_\tau(W') - \frac{4}{W''}\, N_\tau(W'') \right\}^2 \leqslant E(u;\, W',\, W''), \tag{10.2}$$

for an appropriate E arising out of the procedure described in Section 9. A single application of (10.2) with $W' = w'$, $W'' = w''$ leads to an estimate for $B(u;\, \delta,\, w')$ in the manner indicated in Section 9. We now remark, however, that an elaboration of this technique leads to a quantitatively more effective estimate. The additional device is as follows.

The inequality (10.2) enables one to estimate the number of τ (satisfying $d(\tau) \leqslant u$) for which

$$N_\tau(W') \geqslant 8\delta^* n W', \qquad N_\tau(W'') < \delta^* n W'';$$

that is, the number of τ for which (10.1) is true when δ, w' are replaced by $8\delta^*$, W' but false when δ, w' are replaced by δ^*, W''. This estimate (if quantitatively adequate) enables one to deduce from the premise that $B(u;\, \delta^*,\, W'')$ is small (compared to its statistically "expected" value), that $B(u;\, 8\delta^*,\, W')$ is also small (compared to its expected value). The transition from w'' to w' can thus be effected via suitably chosen intermediate widths w_j of type

$$w'' = w_J < w_{J-1} < \cdots < w_1 < w_0 = w'. \tag{10.3}$$

Subject to appropriate hypotheses, (II) tells us that $B(u; 2^{-3J}\delta, w_J)$ is small compared to its expectation, and we can deduce by successive steps for $k = 1, 2, \ldots, J$ (with $\delta^* = 2^{3(k-1-J)}\delta$, $W'' = w_{J-k+1}$, $W' = w_{J-k}$) that $B(u; 2^{3(k-J)}\delta, w_{J-k})$ is also small compared to its expectation. The structure of the bound E in (10.2) is such that, if the system (10.3) is suitably chosen, this technique proves more effective than a single application of (10.2).

The above technique involves (implicitly) the application of (9.19) for various w'' which no longer satisfy the premises of (II). This affects the proof of (10.2) only in that the estimation of the right-hand side of (9.19) requires an additional device in this more general context. (The premises of (II) imply $w'' < \triangle^{1/2}$ and the estimation is simpler in this special case.) The new procedure is as follows.

The expression to be estimated is

$$\int f^2(X)\, dX \leqslant (\max_X f(X)) \int f(X)\, dX, \tag{10.4}$$

where f is defined by (9.16). We know the exact value of the integral on the right-hand side of (10.4), namely,

$$\int f(X)\, dX = \sum_{i=1}^{n} g_i(X)\, dX = nA = n\pi(\tfrac{1}{2}w'')^2, \tag{10.5}$$

so that a suitable bound for $\max f$ will lead to an estimate of the desired kind. But, for any given X, the right-hand side of (9.16) counts the number of i for which P_i lies in the disc $\mathbf{D}^{(X)}$ center X and radius $\frac{1}{2}w''$. Suppose that there are $m = m(X)$ such points P_i, so that $f(X) = m$. If $m \geqslant 3$, we can select three of these m points P_i in $\mathbf{D}^{(X)}$ to constitute a triangle of area not exceeding $\triangle^*(\mathbf{D}^{(X)}; m)$; here we have used the notation of Section 2 (with $\mathbf{K}$ replaced by $\mathbf{D}^{(X)}$ and n replaced by m). But by the extremal property of the distribution $P_1, P_2, \ldots, P_n$ (in the original disc $\mathbf{D}$), no three points P_i can form a triangle of area less than $\triangle(\mathbf{D}; n)$. Thus, again using the notation of Section 2, we have for $m \geqslant 3$,

$$\triangle(\mathbf{D}; n) \leqslant \triangle^*(\mathbf{D}^{(X)}; m) = A\triangle(\mathbf{D}^{(X)}; m) = A\triangle(\mathbf{D}; m), \tag{10.6}$$

where $A = A(\mathbf{D}^{(X)}) = \pi(\frac{1}{2}w'')^2$.

Now, if it is already known that[12]

$$\triangle(\mathbf{D}; t) \ll t^{-\gamma} \qquad \textit{for all } t \geqslant 3, \tag{10.7}$$

[12] The constant implicit in the $\ll$ notation here depends on γ.

it follows from (10.6) that (again writing $\triangle = \triangle(\mathbf{D}; n)$

$$m \ll ((w'')^2 \triangle^{-1})^{1/\gamma} \quad \textit{if} \quad m \geqslant 3.$$

Hence, by (10.4) and (10.5), we deduce

$$\int f^2(X)\, dX \ll n(w'')^2 \max\{1, ((w'')^2 \triangle^{-1})^{1/\gamma}\}. \tag{10.8}$$

We note that the estimate (10.7) is certainly available with $\gamma = 1$. (This is simply the trivial estimate (3.2) discussed in Section 3.)

The use of (10.8) leads to a result of type (10.2) in which the estimate $E(u; W', W'')$ depends on γ. (The larger the value of γ, the better the estimate.) On applying Roth's method with $\gamma = 1$ in (10.2), one obtains a result of type $\triangle(n) \ll n^{-\gamma_1+\epsilon}$ with $\gamma_1 > 1$. We now have (10.7) (and hence (10.2)) at our disposal with $\gamma = \gamma_1 - \epsilon$, and applying the method again (using this strengthened from of (10.2)) one obtains a result of type $\triangle(n) \ll n^{-\gamma_2+\epsilon}$, where $\gamma_2 > \gamma_1$. Clearly this process may be repeated to obtain a sequence of estimates of type $\triangle(n) \ll n^{-\gamma_j+\epsilon}$ ($j = 1, 2, \ldots$) of successively greater precision. In practice the sequence γ_j which arises in this way converges to the root of a quadratic, and this phenomenon accounts for the curious nature of the constants μ, μ' (see (3.6)) featuring in Roth's results.

The application of the estimate (10.2) corresponding to (10.8), in conjunction with the most favorable type of system (10.3), leads (subject to appropriate premises) to the estimate

$$B(u; \delta, w) \ll \delta^{-2} u^2 \triangle^{-3}(u + \triangle^{1-(1/\gamma)} w^{(2/\gamma)-1})\, n^{-1+\epsilon}. \tag{10.9}$$

If δ is small (and w is not unduly small) the rough "expectation" for $B(u; \delta, w)$ is simply the number of τ satisfying $d(\tau) \leqslant u$, and this number has order of magnitude roughly n^2u^2. Note that if $n\Delta$ is not too small, the right-hand side of (10.9) is small compared to this expectation for B, provided δ, u, w lie within appropriate ranges. (This is true even when $\gamma = 1$, the value for which the estimate is at its weakest.)

With regard to results of type (I), we restricted the discussion in Section 9 to a description of the principle which enables one to prove results of that general nature. Methods based on this principle can be adapted in various ways, and there are many possible variants of results of type (I). The effectiveness of each variant for use in conjunction with (10.9) must be measured in relation to the structure of the inequality

(10.9). The result which in conjunction with (10.9) leads to the estimate (3.5) is somewhat elaborate, and we do not state it here; although of the same general nature as (I), the result is not strictly of this type and (in particular) involves the simultaneous consideration of various w' (for a detailed statement and proof of this result, we refer the reader to [4]). In the proof of (3.5) the estimate (10.9) is also applied with various different values of the parameters u and w.

References

1. K. F. Roth, On a problem of Heilbronn, *J. London Math. Soc.* **26** (1951), 198–204.
2. Wolfgang M. Schmidt, On a problem of Heilbronn, *J. London Math. Soc.* **4** (1971/1972), 545–550.
3. K. F. Roth, On a problem of Heilbronn, II, *Proc. London Math. Soc.* **25** (1972), 193–212.
4. K. F. Roth, On a problem of Heilbronn, III, *Proc. London Math. Soc.* **25** (1972), 543–549.
5. K. F. Roth, Estimation of the area of the smallest triangle obtained by selecting three out of n points in a disc of unit area, *AMS Proc. Sympos. Pure Math.* **24** (1973), 251–262.
6. H. L. Montgomery, "Topics in Multiplicative Number Theory," Lecture Notes in Mathematics, Vol. 227, Springer–Verlag, Berlin/New York, 1971.